Session 1
6th May
17:00 h

# UPIMAPI, reCOGnizer and KEGGCharter: three tools for functional annotation

João C. Sequeira[1], Miguel Rocha[1], M. Madalena Alves[1], Andreia F. Salvador[1]

[1] CEB-Centre of Biological Engineering, University of Minho, Braga, Portugal

Omics technologies generate large datasets from which biological information must be extracted by using bioinformatics tools. Although web services provide easier to use interfaces, large datasets are difficult to handle. This is not a limitation of command-line tools and programmatic modules, but these may be challenging to use. In this work, three command-line tools were developed, aimed for speed and automation. The tools are available through Bioconda for Unix systems and were developed in Python 3, making use of multithreading/multiprocessing in computationally demanding steps. UPIMAPI integrates annotation with reference to the UniProt database with automatic retrieval of internal and cross-reference information from other databases (e.g., KEGG, BRENDA and RefSeq) through UniProt's API, accessed with urllib package. The input is a FASTA file containing protein sequences, and the outputs are EXCEL or TSV files containing taxonomic, functional, and cross-reference information. reCOGnizer performs domain-based annotation of protein sequences with CDD, Pfam, NCBIfam, Protein Clusters, TIGRFAM, SMART, COG and KOG as reference databases, and obtains EC numbers and taxonomic assignments per domain identified. The results are outputted in TSV and EXCEL files. KEGGCharter is a command line implementation of KEGG Pathway's mapping service, while also obtaining additional KOs and EC numbers, through the methods available in BioPython for accessing KEGG's API. KEGGCharter takes as input a table (TSV or EXCEL), containing either KEGG IDs, KOs or EC numbers. KEGGCharter represents identified KOs in metabolic maps and includes information on differential gene expression. When data from more than one organism is uploaded, KEGGCharter links function to taxonomic identification, which can be visualized in the maps. Differential expression of genes/proteins can be visualized in metabolic maps, by showing mini heatmaps. UPIMAPI and reCOGnizer are complementary tools, providing functional annotation based on protein sequencing homology and on identification of protein conserved domains, respectively. Both tools retrieve the IDs (KEGG IDs, EC numbers and KOs) necessary to run KEGGCharter. Together, these tools provide a complete characterization and visualization of results, facilitating the interpretation of omics experiments, and requiring minimal bioinformatics expertise.